RESEARCH PAPER

Canonical Workflows to Make Data FAIR

Peter Wittenburg[1†], Alex Hardisty[2], Yann Le Franc[3], Amirpasha Mozaffari[4], Limor Peer[5], Nikolay A. Skvortsov[6], Zhiming Zhao[7] & Alessandro Spinuso[8]

[1]FDO Forum, Gemeindweg 55, 47533 Kleve, Germany

[2]Cardiff University, Cardiff, Wales CF10 3AT, UK

[3]eScienceFactory, 75570 Paris Cedex 12, France

[4]Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[5]Yale University, New Haven, CT 06520, USA

[6]Russian Academy of Sciences, 121351 Moscow, Russia

[7]University of Amsterdam, PO-Box 94323, 1090 GH Amsterdam, The Netherlands

[8]Royal Netherlands Meteorological Institute (KNMI), Utrechtseweg 297, 3731 GA De Bilt, The Netherlands

Keywords: Workflow; Data management; FAIR Principles; Digital Objects

Citation: Wittenburg, P., et al.: Canonical workflows to make data FAIR. Data Intelligence 4(2), 286-305 (2022). doi: 10.1162/dint_a_00132

Received: September 26, 2021; Revised: February 1, 2022; Accepted: February 5, 2022

ABSTRACT

The FAIR principles have been accepted globally as guidelines for improving data-driven science and data management practices, yet the incentives for researchers to change their practices are presently weak. In addition, data-driven science has been slow to embrace workflow technology despite clear evidence of recurring practices. To overcome these challenges, the Canonical Workflow Frameworks for Research (CWFR) initiative suggests a large-scale introduction of self-documenting workflow scripts to automate recurring processes or fragments thereof. This standardised approach, with FAIR Digital Objects as anchors, will be a significant milestone in the transition to FAIR data without adding additional load onto the researchers who stand to benefit most from it. This paper describes the CWFR approach and the activities of the CWFR initiative over the course of the last year or so, highlights several projects that hold promise for the CWFR approaches, including Galaxy, Jupyter Notebook, and RO Crate, and concludes with an assessment of the state of the field and the challenges ahead.

† Corresponding author: Peter Wittenburg (Email: peter.wittenburg@mpcdf.mpg.de; ORCID: 0000-0003-3538-0106).

1. INTRODUCTION

The increasing volumes of data on the one hand and the request for proper documentation to enable repeatability and reproducibility on the other require a change in the practices of data science laboratories [1] and a change in the role of researchers. The FAIR principles [2, 3, 4, 5] have been developed as guidelines for changing data practices. The core message behind the four principles is "machine actionability", which means that when machines find a value in some data structure, they know how to behave—how to interpret the value based on some widely agreed semantic definitions and/or on an explicit relation with a procedure so that computations can be carried out. This formulation clearly indicates a direction toward increased automatic processing of the huge masses of data being created by contemporary science.

The role of researchers in the field of data science will change as well. Currently, data and reporting standards are not sufficiently well-developed. Data scientists/researchers often rely on colleagues they know and trust, who are creating suitable and reliable data. Given the increasing number of data labs and the productivity of next-generation experimental methods, the status quo is unsustainable and will only prepare a small subset of data for reuse over the long-term. We claim that automated registries offering data with detailed metadata description must evolve further, allowing researchers to specify query profiles and to activate agents on their behalf to carry out searches and locate, negotiate terms of use, import and transform suitable data from labs worldwide.

These changes seem to be inevitable. However, they will take time. This impression is supported by a recent study that offers insights into the work and plans of about 75 current research infrastructures initiatives [1]. We refer to two paradoxes mentioned in that study to be addressed by the concept of Canonical Workflow Frameworks for Research (CWFR):

- **Data science has been slow to embrace workflow technology despite clear evidence of recurring practices and decades of funding**: There is clear evidence of many recurring patterns in data science carried out across the disciplines. However, workflow technology is not yet sufficiently widely embedded in routine practices despite decades of funding technological developments.
- **Research practices are slow to fully integrate FAIR principles despite growing awareness of their utility**: In laboratories and centres creating, managing, and processing data, awareness is growing of the relevance of the FAIR principles, but data science has yet to integrate them into its general practice and methods. FAIRness (i.e., compliance with FAIR principles) is often only brought up towards the end of a project, when publishing final results. This contrasts with the principle of Open Science (FAIR) by Design, as described by the Board on Research Data and Information (BRDI)[①], which expects FAIRness throughout the research lifecycle and therefore opening the mass of digital objects in a FAIR compliant way[②].

① https://www.nationalacademies.org/our-work/toward-an-open-science-enterprise

② It is the current practice amongst many collaborators to exchange masses of data at early stages; however, this is currently being done at the file level without further information.

The above indicated paradoxes illustrate the challenges of generating a domain of FAIR compliant Digital Objects (FDO) [6] that are necessary for flourishing data science. The gap between practices on the one hand and workflow technology and FAIR compliance on the other hand is great and begging for new approaches. Researchers are slow, even reluctant to incorporate FAIR in their daily practice since this would require considerable adaptations and re-developments—which would reduce productivity for some period. They even may not know how to do this. Using current workflow technology would imply dependency on software development experts which are scarce or on the need to learn appropriate software skills. These options are not attractive for most researchers.

This paper describes the goals and activities of the CWFR initiative. The initiative has twin goals: To bring experts together to better understand what is currently being done with respect to data in the different research labs and to examine how to improve data practices stepwise by introducing canonical workflows based on prefabricated and harmonised components that produce FDO compliant and thus interoperable documentation. Such frameworks will be accepted by researchers and help change practices only if they do not add to or seriously disrupt their work. Therefore, the CWFR initiative aims to offer a platform where experts from research infrastructures can exchange knowledge and solutions, extract a realistic account of the state of workflow technology in various labs (being aware that this may be a limited view), and urge change towards more FAIRness.

1.1 Workflows, FDOs, and the CWFR Initiative

We begin with a brief explanation of key terms. The newly developed concept of FDO combines the already widely used concept of Digital Objects [6, 7, 8] with the FAIR principles [2, 3, 4, 5] i.e., ensuring that all aspects of DOs are FAIR compliant and thus machine actionable. Rather like POSIX file systems, Digital Objects is a simple concept to abstract and unify the access to all different data types and data organisations. DOs basically are a structured bit-sequence encoding content and stored in some repositories, with an assigned globally unique, resolvable and persistent identifier (PID) and rich metadata to describe them. DOs represent the highest level of abstraction, since the bit-sequence can cover any kind of content (data, metadata, software, assertions, etc.), since the assigned PID should refer persistently to all information aspects that are needed to find, access, interpret and reuse its bit-sequence and since its "type" as part of its metadata will enable linking it to procedures (methods, operations). It should be noted that the FAIR principles do not make statements about "openness" of data. In contrast, due to their binding capability[9] FDOs allow applications to build in security measures controlling access to content. However, despite its relevance security aspects are not touched further in the present paper and in the present special CWFR issue.

It should be said here that programmatic scripts providing fragments of more comprehensive workflows are already widely used in system management and in various scientific domains, but they in general cover workflow tasks that are not the subject of variations. Developments now aim to be more comprehensive,

[9] FDOs bind all information relevant for processing together via its PID.

more flexible and to tackle recurrences by using new kinds of technological frameworks (e.g., Galaxy, Jupyter, etc.). There are also examples of well-used comprehensive workflow systems created to be used by laypersons [10] and ready-to-use systems generated with the help of Galaxy [11], for example. The above cited report [1], which is based on a broad analysis, clearly indicates however, that despite these exceptions more comprehensive workflows tackling the possible recurrences are widely lacking. The situation in industry is different, since recurring patterns in industry in general are much more rigid.

Of course, the spectrum of workflow approaches applied in labs creating, managing, and processing data is already wide and will continue to be so, as indicated in Figure 1.

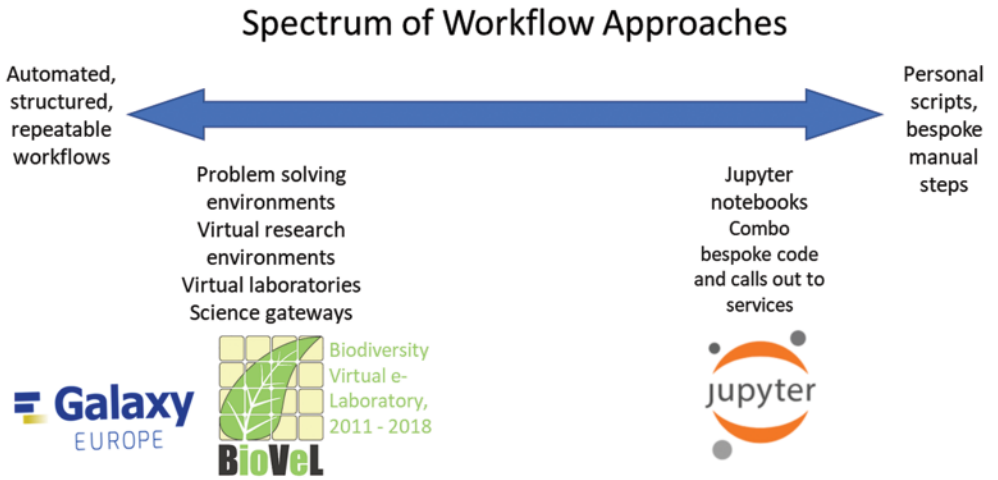


Figure 1. Spectrum of workflow approaches in the research domain.

Some experts will prefer to write personal scripts or use bespoke manual steps, while others will prefer to rely on orchestrating their automatic workflows with readymade components which are taken from open software libraries[4] or libraries of accessible commercial platforms[5]. As indicated, Galaxy [11] and BioVeL [12] can be seen as relatively recent solutions that support the different approaches. Galaxy and Jupyter are quite generic, whereas BioVeL was an example targeted towards a specific domain and community. While Galaxy and BioVeL are intended to offer components that can be reused to orchestrate workflows, Jupyter Notebook [13] is a general-purpose interpretative framework that is meant not only for quickly developing and testing scripts but also acting as a mechanism for documenting and publishing code and results. However, as we will show, each is still quite out of reach for the general researcher lacking skills for programming and using remote computing facilitates.

[4] It should be noted that personal scripts etc. can be seen as first steps to more comprehensive solutions.
[5] Google Earth Engine, for example.

When workflow frameworks are designed to offer a rich component landscape, generic methods for describing the interactions between the components of the workflow and with the workflow management must be provided. Often, proprietary workflow languages and database solutions are being used internally. These were not designed with FAIR compliance in mind, but primarily to support the efficient operation of the workflow management system (WfMS). Various export formats often enable the exchange of documents or artefacts being created at the end. When all tools and frameworks, however, support the creation and usage of FDOs as first-class citizens on the Internet, a much higher degree of interoperability would be achieved, independent of changing technological fashions. Stated another way, when sufficient state information is available for the workflow, its inputs and outputs, and the circumstances of its execution, it becomes fully computationally reproducible [14] and reusable in the same or any other compatible WfMS framework.

Workflows can also be distinguished along the user involvement dimension. Many workflow frameworks are optimised towards an automatic execution of orchestrated workflows without user intervention. In general, these are called "computational workflows". Some workflow frameworks support user involvement in so far as they allow users to intervene and steer further processing. Workflow execution is interrupted to carry out some manual activity and then to continue the execution after some period. This can also imply parallel execution of different tracks and thus asynchronicity. An example of such asynchronous parallelism can be observed in experiments with human subjects. In general, researchers must ask for individual consent and ethical permissions which take time to be evaluated and granted. In parallel, the researchers may continue with some tests to see whether the experimental paradigm and the setting of parameters meet the study expectations. In addition, progress with each individual subject can proceed at a different pace. In all these cases, proper documentation of the state of the workflow is essential.

The Group of European Data Experts (GEDE)[©], a loose collaboration between data professionals and experts from research infrastructures, has discussed these paradoxes and has suggested launching the present "Canonical Workflow Framework for Research" initiative, based on several major convictions:

- We cannot expect that every data scientist will become a professional in developing efficient workflow tools.
- Current workflow technology is still quite removed from the daily practices in many data labs. It must be brought closer to researchers' practices by introducing workflow frameworks and libraries of canonical components that can be easily learned and easily reused.
- Such workflow frameworks, which should be sufficiently attractive for researchers to adopt (ease of use, wide range of familiar tools) must support FAIRness internally i.e., the researcher should not be bothered by the demands of meeting FAIR criteria for different objects (e.g., data, software, and configurations) involved in the workflow. These should be handled automatically.

[©] The GEDE group (RDA GEDE: https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda) has recently become part of the FDO Forum.

- The best way to achieve FAIRness in workflow technologies is to request that those canonical workflow components that are available to be integrated into workflow solutions should support the concept of FDO [9], meaning that these components should be capable of both consuming and producing FDOs.

The deep insights gained by analysing several advanced research infrastructure projects and plans by Jeffery et al. [1] also contributed to the launch of the CWFR initiative.

2. CANONICAL COMPONENTS FOR WORKFLOWS

We described the aspect of recurring patterns and the inefficiencies in data-driven research on the one hand, and the aspect of current workflow technologies and missing support for FAIR practices on the other hand. The question that must be addressed is how to overcome these deficiencies: How can communities carrying out data-driven research derive greater benefit from (i.e., be encouraged to make wider and better use of) workflow technologies for their work? What practical functionalities can workflow management systems introduce for better supporting FAIR principles and practices throughout the research lifecycle—from first hypothesis and inception of a project to its conclusion with publication and archiving of open FAIR results?

While a change of working practices is a long-term goal, it can only be achieved incrementally and will take time. Adherence to the FAIR principles, which means machine actionability, has been identified as crucial to reducing inefficiencies. The FDO technical framework [15] that will help with turning FAIR into practice has been specified already, and further work is underway sponsored by the FAIR Digital Object Forum (FDO Forum) [16]. However, FAIRness cannot be achieved by human actions alone. It would cost too many human resources and be prone to omissions and postponements.

Introducing canonical workflows into this journey can address both paradoxes described. By ‘canonical workflows’ we mean workflows composed of somewhat standardised research lifecycle components that produce FAIR compliant outputs or artefacts as a matter of course.

First, adopting canonical workflows can help with automation of recurring processes in data science. Second, because of their characteristics, canonical workflows ensure that intermediate component outputs automatically support the FAIR principles and practices from beginning to end of the work. These are features researchers can easily recognise as benefitting their work, thus making them more willing to take the step. Workflows and WfMS conforming to and supporting the requirements of FAIR and the technical framework(s) and specifications of FAIR Digital Objects motivate the concept of a Canonical Workflow Framework for Research (hereafter, CWFR), which makes use of FDOs as the underlying documentation and interoperability standards. We claim that such a framework can be of central significance to improving FAIRness for many communities.

Our analysis of use cases presented at a workshop [17] and a deep analysis of a large number of research infrastructure projects [1] in 2020 showed that it is indeed possible in many research domains to identify recurring patterns and canonical components being repeatedly applied throughout all stages of typical research processes. Such canonical components are "actions" that are understandable to the researchers, such as "carry out an ethical review", "select appropriate subjects for an experiment", "perform an experiment", "process some data", "perform a machine learning step", etc.

For many of these actions, common practices and software tools already exist. Yet, these tools have often been developed as stand-alone applications offering some needed functionality, with little thought to compatibility with other tools. These tools produce outputs which users need to store and transform manually so that they can be used during subsequent actions, i.e., outputs/inputs are typically not at all integrated in a seamless manner, frequently resulting in fragmented and potentially irreproducible sequences that mix manual and machine-based steps. Typically, there is no systematic documentation of what has been done. There is very little support for FAIR among existing tools. To integrate these components into a CWFR, it will be necessary to:

(1) Put them into shareable libraries so that they can be integrated during the composition phase,
(2) Offer a workflow framework that embeds them and creates the FAIR compliant documentation or artefacts for each action, and
(3) Develop mechanisms that provide wherever feasible the import and export transformations to remove interoperability incompatibilities between components. In many cases it will be necessary to develop wrappers to make the existing components "CWFR ready".

It is obvious that this will be an incremental process that can only be comprehensibly and comprehensively enacted if many researchers see the need and advantages of supporting automation and of creating a FAIR domain of exchangeable digital objects.

These ideas are well-known in the workflow technology developer community, yet with some exceptions the solutions have not yet been addressed from the point of view of the researchers and this is essential for achieving breakthroughs. There are some solutions such as, for example, Weblicht [18], that were designed to help the naive computer users. Weblicht allows linguists to orchestrate typical natural language processing (NLP) processes, e.g., "identify named entities" in a text. The user enters a text in a specific language which is identified by the workflow machine and then appropriate operations are offered from a library starting with part of speech tagging. After each step the machine offers the remaining valid choices of components, i.e., the machine knows about the previous actions which are documented in the metadata and therefore knows which operations are possible for the given output at that step. All interaction is thus at a level that linguists easily understand. However, Weblicht, like other comparable tools, only addresses a specific set of applications. The chosen metadata format is specific for this application domain and FAIR compliance was not a primary objective, i.e., all was contained in one specific DO.

Requesting the participation of many research fields and still achieving interoperability requires a clear and commonly acceptable definition of "the glue" that all components adhere to. This implies that the core FDO that records the state and context of a workflow in a comprehensive manner must be very well specified. To make it fully FAIR compliant and thus machine-actionable would also require making use of strict typing (i.e., rules about the properties (attributes) of information that must be preserved) and registration of component types in a well-known registry. There have been attempts to use strict typing as a means of supporting interaction between a variety of components. Apache UIMA [19], originating from IBM Research, is a specification based on strong typing, i.e., every component needs to define in detail what its import and export types are, allowing others to make use of information if necessary. UIMA offers a scheme with global type specifications and slots for components to add their specifications. The OPC Unified Architecture [20] standardisation in industrial automation is based on the willingness of all manufacturers of production machines to define and register the input and output variables in addition to the formats. Research Objects (RO) and RO Crates [21, 22] also offer a framework with specified attributes to describe the workflow result package.

Such standards and best practices must be carefully analysed respecting the views of different stakeholders to design a set of specifications ready to be used within CWFR. In the research domain, CWFR would imply that developers start anticipating the need for embedding tools in workflows and define and register the attributes being used for input and output. This is a significant expenditure of effort needed to overcome the deficiencies described at the beginning of this article. However, once critical mass is reached, the improved interoperability and reproducibility will greatly improve the productivity and reliability of data-driven research. It will empower domain researchers to innovate within their specialism in an agile manner, without being dependent on assistance from workflow specialists.

3. CONTRIBUTIONS TO THE CWFR INITIATIVE

The CWFR initiative organised five working meetings from November 2020 until July 2021 to intensively discuss the state of workflow applications and technology and how to overcome the deficiencies of current practices. The meetings offered a rich mix of use case overview, in-depth presentations, and community discussion. It was the intention of the CWFR initiative to act first as a platform to understand how workflow technology is being used in different scientific institutions and then to discuss ways by which solutions could evolve to address the needs of researchers that are still far away from adopting workflow solutions. This paper therefore offers an overview of what has been presented and what kind of conclusions can be drawn from the presentations[7]. Many presentations in these meetings have then been transformed into papers in this special issue.

[7] We note that this paper mainly reflects on the meetings and the discussions which have taken place thus far as part of the CWFR initiative and which can be found as fuller contributions to this special issue. The discussions during the workshops reported below and some paper contributions indicate that currently much more training courses on workflow technologies are given. They range from generic extensions of Python to Jupiter courses which are well perceived by many young students to training courses about complex systems such as Galaxy. In this special issue and this paper these efforts are not in the focus although they are very important for improved take-up in the future.

3.1 2020 Working Meetings: Use Case Overview

At the first two meetings, held on 16th and 20th November 2020 [17], sixteen researchers and developers from different disciplines were invited to present concepts and solutions. This group of presenters is not necessarily representative of practices at all data labs in universities and research institutions as indicated in the introduction. The meetings confirmed that an increasing number of research groups are starting to experiment with workflow frameworks to cover some selected recurring patterns of sequences of actions. It also became apparent that new frameworks and formats are being invented with the intention of making it simpler for the individual researcher. As in all sectors of technology we can observe another fragmentation which is driven by technological dynamics and the need for more efficiency in data-driven work. The growing transformations and new expressions of the workflow landscape brought about by the interactions of WfMS developers with specific end-user communities, where new varieties of standard WfMS forms and functions are made into new varieties superseding the old, is exciting. It reveals the implementation of many new ideas. Yet, in this dynamic landscape, interoperability is an increasing challenge, and FAIRness does not yet seem to be a primary concern.

Interpretative frameworks such as Jupyter Notebook seem to be attractive to many groups and departments because of training courses for early career researchers and inclusion in some undergraduate student curricula. This has a positive effect. However, the assumption that the broad community of researchers will become developers of reusable code is questionable.

Another conclusion from the different use cases presented in these meetings is that the commonalities with respect to process patterns, reusable components, and interoperability standards are at first glance not as evident as one might hope. More intensive analysis work must be carried out within disciplines to identify commonalities. It was obvious, however, that the introduction of a structured document with strongly typed parameters that captures all needed information to describe the state of a workflow in a comprehensive manner could serve as an interoperability 'glue.' Therefore, the concept of FAIR Digital Objects, which should anticipate best practices such as Research Object Crates, UIMA (Unstructured Information Management Architecture) and OPC Unified Architecture seems to be a worthwhile approach to pursue.

3.2 March 2021 Meeting: Galaxy

The March CWFR working meeting was devoted to discussing the potential and possible limitations of the Galaxy workflow framework. The Galaxy framework development was started about 15 years ago in bioinformatics to increase the efficiency of data science in biology and to address the lack of reproducibility of studies. Galaxy was intended to support "typical computational workflows" in biomedical sciences and therefore it allows experts to include existing tools by providing wrappers, describing them by metadata and integrating them into tool libraries (Figure 2).

Figure 2a. Wrapping standalone tools to integrate in Galaxy workflows.

Figure 2b. Galaxy allows integration of existing tools by developing wrappers and provides several standard adapters to import and export commonly used data formats in the biomedical domain.

During composition, users can select tools from libraries and combine them to form workflows. As in other frameworks, the output-input relationships need to be solved, i.e., transformers that turn output parameters into appropriate input parameters of the subsequent action must be provided. The Galaxy developers provide modules that can do this transformation between a set of well-known tools. For other tools, new transformers would have to be developed. All process information, the workflow script, the state of the workflow, references to information being used, etc. are maintained in a relational database. For documentation purposes, reports can be generated in a variety of formats, which range from PDF documents that can be stored in Zenodo up to structured information using description standards such as RO-Crate. Where feasible, Galaxy makes use of the EDAM ontology [23] which has been developed by the bioinformatics community.

This brief description indicates that the concept behind Galaxy matches closely with the ideas behind CWFR and it is not surprising that Galaxy is being used widely in the biomedical community, with some attempts at use beyond that community. Some researchers create and execute new workflows, but many reuse prefabricated workflows for a variety of computations. Galaxy contributes greatly to closing the gap between workflow technology and researchers. However, we see a few limitations and can draw some major lessons:

- Galaxy is supporting only synchronous workflows at this moment, i.e., asynchronous operations which occur when processes are, for example, waiting on human intervention or on the end of the process to be executed in parallel, are not (yet) supported.
- The workflow and state documentation is included in an optimised relational database which is not designed to enable sharing (and therefore switching between environments, for example). Currently, comprehensive reports are only provided at the end of workflows.

- The integration of tools is not a trivial task and requires experts to create the required wrappers.
- Creating the output-input transformations is also not trivial and requires a general deep understanding of the environment.
- When tool integration and parameter transformations have been solved it is comparatively easy for researchers to orchestrate and execute workflows.

In sum, from a CWFR point of view, we see Galaxy as a major step to close the gap between technology and researchers. For the integration of tools, the implied input-output parameter transformation, and the documentation of workflows and their states, it seems that we have not yet achieved a satisfying state.

3.3 April 2021 RDA Workflow Meeting: Community Discussion

The discussion at the 17th Plenary Meeting of the Research Data Alliance® indicated that while in many cases it will be challenging to embed research work in automatic workflows, an increasing number of examples suggests that much data science work—from creation to analysis—does or will include recurring sequences of steps that can be automated. It was generally agreed that well-designed workflow frameworks for proper documentation of what has been done is important, for example, to increase reproducibility and improve FAIR compliance.

Inventing mechanisms to make it easier to create proper input-output relationships between subsequent processing steps was also addressed. In addition, for requesting strict typing of attributes by defining and registering them in ontologies, it was suggested to develop a standard for defining "units" as is done for example by i-Adopt [24, 25]. Making relationships explicit by registering them, as suggested by the initiative on Semantic Mapping Framework (SEMAF), would make relationships sharable across frameworks [26].

3.4 June 2021 Meeting: Jupyter Notebook

At the June meeting®, the potential and possible limitations of the flexible Jupyter Notebook framework were discussed. Jupyter Notebook has become increasingly attractive to young data scientists. Two of the major reasons for this are its interpretative nature, allowing fast prototyping and its ease of use with Python programming, which is now taught in many university courses on data science across disciplines. Therefore, education programmes would do well to extend training to include Jupyter and the speakers reported promising results. However, training needs to include the writing of proper documentation, using facilities such as Markdown [27], to make it possible for future users to understand the code.

Jupyter is a generic software framework and therefore, can deal with asynchronous activities and user interactions. However, in contrast to frameworks like Galaxy, it requires more basic software development skills. Jupyter enables the integration of existing software by providing wrappers that need to be developed,

® https://www.rd-alliance.org/canonical-workflow-frameworks-research-cwfr
® https://osf.io/ywcq4/

including the transformation of parameters between processing actions. Since it supports generic programming languages, Jupyter would allow building a reusable workflow framework such as sketched by CWFR. However, the setup of software libraries, the choice for metadata descriptions, the processing of collections, the generation of appropriate and FAIR compliant documentation with the help of FDOs, etc., would have to be developed by a community interested to extend Jupyter [28].

As a generic framework, Jupyter poses few restrictions but requires significant intervention by professional software developers in order to become an easy-to-use framework for canonical components interacting via FDOs. It seems to be a useful framework for researchers with modest programming skills to implement workflows that are being used personally or in a small team of known users. However, for complex tasks such as executing workflows on High-Performance Computing (HPC) machines, or for use by a large community of potentially unknown users, building blocks would have to be provided by experts.

It was agreed that (1) it cannot be assumed that all researchers become workflow programmers and (2) in general it should not be expected that researchers trained in Python will go on to become professional software developers capable of tackling more complex tasks and producing robust, reliable, sharable and maintainable code.

3.5 July 2021 Meeting: RO Crate

Research Objects (RO) [21] is a concept for packaging all resources required to execute a workflow and thus to achieve repeatability and increase reproducibility. Resources of any type can be integrated as objects or as references pointing to remote objects. It is known that this is one of the big challenges in data-driven science that needs to be tackled and therefore suggestions for containers are of great importance. The concept of RO emerged from biomedical sciences but has also been taken up in various other sciences (digital humanities, earth sciences, regulatory science, . . .). It is implemented by providing a community specification called RO Crate [22] which is based on JSON-LD and where possible makes use of vocabularies anchored in "schema.org"[8]. RO Crate offers examples with selected category sets for a variety of objects. Discipline-specific metadata can be associated with RO Crates which is important since in many cases communities have already specified the schemas and vocabularies they are using.

RO Crate is a self-described container which contains a set of data files and is described by a single metadata document describing the resulting object and its individual components[9]. Each of these can be described by a set of usual metadata categories taken from the "schema.org" registry and own extensions. For distribution purposes container formats such as ZIP and BagIt can be used. Examples are indicating that semantic extensions are possible to capture specific metadata which is being used in different research communities.

[8] "schema.org" is a registry for schemas and semantic categories relevant for describing webpages. It is supported by many companies and is owned by Google.

[9] In RDA Research Collection and FDO terminology this is called a collection.

As a container, RO Crate offers ways to capture, aggregate and document the essentials of workflows independently of the workflow framework chosen (CWL, Galaxy, Jupyter, WfExS, etc.) in a way that allows them to be stored as workflow objects that can be exchanged and executed at another location. This approach supports reproducibility and repeatability of computational workflows. RO Crate includes provenance information to trace the genesis of the resulting output. Some efforts are on the way to use RO Crate to generate workflow building blocks such as the WfExS presented during the workshop with the clear intention to exchange them between workflow frameworks.

Similar RO Crate flavours have been created for repositories, data citations, 4-dimensional objects etc. Recently, the RO Crate team started to extend its formal framework such that it is compliant to the FDO Framework specifications. Since RO Crate is basically a flexible metadata framework the steps to make it FDO compliant are straightforward, independent of whether the Linked Data (including the FAIR Signposting approach[©]) or the Digital Object (including Handles/DOIs) approach is being chosen. RO Crate describes the structure and metadata of the Research Object using defined and registered vocabularies. Depending on the approach, different PIDs can be added. The content of a collection is a structured container that is being referenced in the metadata. The RO Crate metadata files are FAIR Digital Objects since PIDs could be registered to point to them and since their type can be linked to processors.

In sum, one can state that RO Crate is a well-designed and elaborate metadata solution for different types of objects including aggregation of objects. It relies on JSON-LD syntax and making use of vocabularies registered in "schema.org". It can be used as a particular FDO packaging all the individual resources independently of their types that are being used within workflows as FDOs. For more information, please see: https://osf.io/v8xjz/.

4. CONCLUSIONS

Since 2016, the FAIR principles have been accepted globally as guidelines for improving data science and management practices. At the same time, it has become clear that without major steps in technology support, practices will not and cannot change. Researchers who need to produce high-quality research results within short time-periods are primarily interested in functionality extensions. Implementing or following standards that show positive productivity and/or reproducibility effects for subsequent researchers following after the work of the original researchers is much less justifiable and interesting. The incentives for change are presently weak.

The CWFR working meetings, the call for papers on canonical workflows leading to the present special issue, as well as other events, indicate that an increasing number of institutions are looking at and experimenting with workflow technology in the wide context of whole-research lifecycles and processes and are organising training events. In most cases such workflow projects are undertaken where IT experts can lead the development efforts. This is especially true when generic frameworks such as Jupyter Notebooks

[©] https://signposting.org/FAIR/

are being used beyond comparatively simple pipelines which imply few restrictions but require proper software development skills. On the other side of the technology spectrum are workflow applications such as for example Weblicht, Galaxy and YARD [29] that address specific tasks, albeit with some restrictions. The example of Galaxy shows that, although it can be characterised as general-purpose across biomedical sciences, it is still an effort to integrate existing tools and to solve the import-export transformation challenges which requires expert knowledge, even in that single domain. For ready-made processes where the tool integration and the import-export challenges have been solved, the use of Galaxy indicates that this can lead to considerable uptake.

Therefore, we can conclude that the concept of canonical workflows for research framework is a promising way to bring about more FAIR data practices. We can also conclude that this endeavour entails significant work: It requires analysing processes in all disciplines to identify recurring patterns and their variations and to make the tools already existing, and those that will be developed in future, "CWFR ready" for integration. This work will likely involve tool adaptation or the development of wrappers. In the long run, it will also need to address the import-export challenges, which can best be solved by an incremental specification of the parameters for each of the integrated tools--which implies registering their structure, constraints, and semantics in a trusted registry. As has been shown for Apache UIMA and OPC UA this is not a trivial task. However, with a behavioural change whereby all tool builders are required or incentivised to specify their input and output parameters, the effort is quite feasible. An increasing set of modules could be provided that would automatically carry out the necessary transformations. Where this would not help it should become simpler to specify transformations.

Our assessment, based on the many contributions and discussions at our meetings, is that we can be sure that new workflow technologies will be developed with the intent of making it easier for researchers to apply them in their daily work. This innovation will be ongoing. The technologies surveyed here have their own way of representing the workflow and in particular, the state of workflow processes. The CWFR initiative suggests a standardised approach, including an agreement on an interchangeable documentation format with FAIR Digital Objects as anchors. Furthermore, the scientific community also needs to come to an agreement on a standardised state exchange format so that after every action, a comprehensive record is being stored, including all relevant information about the preceding steps. The work of the RO-Crate initiative and the methods used in Apache UIMA and OPC UA and other harmonisation initiatives can inform and generate suggestions in this regard. Finally, the scientific community must do more to develop and to recognize the value of data and information professionals who can support these efforts on an ongoing basis.

Discussion in the CWFR meetings has shed light on the excellent work by various initiatives, but there is still much work needed to implement the CWFR concept and thus to reduce the gap between researcher practices on the one hand and the lack of FAIRness and efficiency on the other hand. First adaptations can be seen for example by projects adopting the emerging FDO standard in their implementations, and plans to improve FAIR compliance and thus increasing interoperability and reuse. Frameworks such as Galaxy and UIMA obviously move already towards registries of harmonized canonical components. These libraries

of harmonized components are reused mainly within specific scientific domains yet these efforts sharing similar semantic spaces and analytic tools, require considerable IT-expert support, are using their own structures and semantics and are limited to computational workflows which do not cover all steps in data science. More effort towards a widely accepted documentation standard based on FDOs should be invested based on the experience already gained.

AUTHORS CONTRIBUTIONS

Yann Le Franc (ylefranc@esciencefactory.com) is co-chair of the SEM working group of the FDO Forum, co-organised the CWFR workshop about RO-Crate and wrote the corresponding section. All other authors are co-chairs of the CWFR working group of the FDO Forum and thus responsible for the planning and organisation of the CWFR work and meetings and thus are co-authors of the paper.

REFERENCES

[1] Jeffery, K., et al.: Not ready for convergence in data infrastructures. Data Intelligence 3(1), 116–135 (2021)
[2] Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016)
[3] Mons, B., et al.: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. Information Services & Use 37(1), 49–56 (2017)
[4] Mons, B., et al.: The FAIR principles: First generation implementation choices and challenges. Data Intelligence 2(1–2), 1–9 (2020)
[5] Jacobsen, A., et al: FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(1–2), 10–29 (2020)
[6] Kahn, R., Wilensky, R.: A framework for distributed digital object services. International Journal on Digital Libraries 6, 115–123 (2006)
[7] Kallinikos, J., Aaltonen, A., Marton, A.: The ambivalent ontology of digital artifacts. MIS Quarterly 37(2), 357–370 (2013)
[8] Hui, Y.: On the existence of digital objects. University of Minnesota Press, Minneapolis (2016)
[9] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publications 8(2), Article No. 21 (2020)
[10] Ferrucci, D., Lally, A.: UIMA: An architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 10(3–4), 327–348 (2004)
[11] Afgan, E., et al.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, W537–W544 (2018)
[12] Hardisty, A.R., et al:. BioVeL: A virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecology 16, 49 (2016)
[13] Kluyver, T., et al.: Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90. IOS Press, Amsterdam (2016)
[14] National Academies of Sciences, Engineering, and Medicine, et al.: Understanding reproducibility and replicability, reproducibility and replicability in science. National Academies Press, Washington, DC (2019)

[15] Research Data Alliance Group of European Experts (RDA-GEDE). (2019). FAIR digital object framework version 1.02, November 2019. FDOF Technical Implementation Guideline. Available at: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF. Accessed 11 August 2021
[16] FAIR Digital Objects Forum (n.d.). Available at: https://fairdo.org/. Accessed 11 August 2021
[17] CWFR workshop. Available at: https://osf.io/9ut4p/. Accessed 11 August 2021
[18] WebLichtWiki. (n.d.). Available at: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page. Accessed 11 August 2021
[19] Apache UIMA. Available at: https://uima.apache.org/. Accessed 11 August 2021
[20] OPC Unified Architecture. Available at: https://opcfoundation.org/about/opc-technologies/opc-ua/. Accessed 11 August 2021
[21] Bechhofer, S., et al.: Research objects: Towards exchange and reuse of digital knowledge. Nature Precedings (2010). Available at: https://www.nature.com/articles/npre.2010.4626.1.pdf. Accessed 11 August 2021
[22] Soiland-Reyes, S., et al.: Packaging research artefacts with ro-crate. arXiv preprint arXiv:2108.06503 (2021)
[23] Ison, J., et al.: EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29(10), 1325–1332 (2013)
[24] Magagna, B., et.al.: InteroperAble descriptions of observable property terminology WG (I-ADOPT WG). Available at: https://www.rd-alliance.org/groups/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg. Accessed 11 August 2021
[25] Gil, Y., Ratnakar, V., Fritz, C.: Assisting scientists with complex data analysis tasks through semantic workflows. In: AAAI Fall Symposium Series on Proactive Assistant Agents, pp. 14–19 (2010)
[26] Broeder, D., et al.: SEMAF: A proposal for a flexible semantic mapping framework. Available at: https://zenodo.org/record/4651421#.YRplHt9CTb0. Accessed 11 August 2021
[27] Gruber, J.: Markdown. Available at: https://daringfireball.net/projects/markdown/. Accessed 11 August 2021
[28] Rule, A., et al.: Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLOS Computational Biology 15(7), e1007007 (2019)
[29] Peer, L., Dull, J.: YARD: A tool for curating research outputs. Data Science Journal 19(1), 28 (2020)

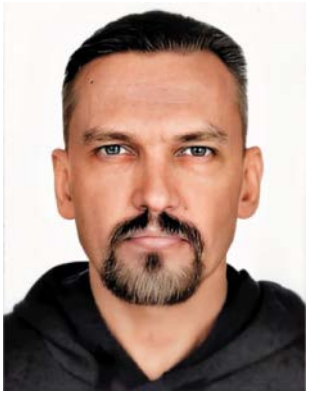
AUTHOR BIOGRAPHY

Peter Wittenburg has a background in electrical engineering, has been working as Technical Director at the Max Planck Institute for Psycholinguistics for many years and acted as member of the IT Advisory board of the president of the Max Planck Society (MPS). The Max Planck Institute was from the beginning focusing on digital technologies to understand the functioning of the brain with respect to language processing. The institutes need for getting access to data from other institutes to feed the stochastic engines they applied rather early led him to become an expert in building data/research infrastructures. He was responsible for the technological aspects of three large international and European research infrastructures: DOBES, CLARIN and EUDAT. In this function he understood that data work across silos is highly inefficient and that harmonisation and standardisation is required to improve the situation. This was the reason that he co-founded the Research Data Alliance in 2013 and the FAIR Digital Object Forum in 2019.
ORCID: 0000-0003-3538-0106

Alex Hardisty was before his recent retirement Director of Informatics Projects in the School of Computer Science and Informatics, Cardiff University, UK. He is interested in bio/geodiversity informatics, the engineering of large-scale distributed information systems for data management and processing, virtual research environments and socio-technical issues of new technology adoption. Alex is a technical architect. Before his retirement he was leading DiSSCo technical work on open Digital Specimens (openDS), Minimum Information about Digital Specimens/Collections (MIDS/MICS) and exploiting machine-actionable FAIR Digital Objects. Alex was co-chairing the CWFR working group of the FDO Forum and was a member of the FDO Forum's Technical Specification and Implementation (TSIG) working group and the FDO Forum Steering Committee.
ORCID: 0000-0002-0767-4310

Amirpasha Mozaffari is a postdoctoral researcher of the group on Earth System Data Exploration (ESDE) at the Jülich Supercomputing Centre (JSC). He is trained as a geoscientist and recently defended his Ph.D. in Computational Geohydrophysics from RWTH Aachen. He is active in the field of data management, workflow design and FAIR data practices. He is co-chair of the Canonical Workflow Framework for Research in the Fair Digital Object Forum.
ORCID: 0000-0001-6719-0425

Alessandro Spinuso is a researcher at the R&D Observations and Data Technology division of the Royal Netherlands Meteorological Institute (KNMI). He earned his Ph.D. in Computer Science at the University of Edinburgh, UK in 2017. At KNMI, he covers the roles of Researcher and Product Owner within an Agile R&D team developing Provenance-aware Data Analysis services. His main research interest is the management and the exploitation of provenance information in the context of user controlled computational environments, providing notebooks and workflow systems for data-intensive analysis. He is involved in several international initiatives focusing on the development of e-science infrastructures for Earth Science research in Europe (EPOS, ENVRIFair, IS-ENES3, DARE, C3S). More recently, he is an invited expert to the IPCC TG-Data. A working group has been dedicated to the FAIR management of the data and methods that will be published in the next IPCC reports.
ORCID: 0000-0002-0077-8491



Nikolay Skvortsov has been affiliated with the Institute of Informatics Problems, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia. His general research interests are ontological and conceptual modeling of research domains and data semantic interoperability issues. In recent years Nikolay Skvortsov investigates requirements for the reuse of data, research methods, and processes in research communities primarily using examples of problem development and solving in astronomical research domains.
ORCID: 0000-0003-3207-4955

Limor Peer, Ph.D., is Associate Director for Research and Strategic Initiatives at the Institution for Social and Policy Studies (ISPS), Yale University. Limor works on research transparency and reproducibility and is especially interested in the connection between generating and preserving scientific knowledge. Limor created the ISPS Data Archive, a digital repository for research produced by scholars affiliated with ISPS with a focus on experimental design and methods. She led the project to develop YARD, the Yale Application for Research Data, a workflow tool for reviewing and enhancing research outputs. Limor is co-founder of the CURE (Curation for Reproducibility) Consortium of social science data archives. She co-chairs the CURE-FAIR working group at the Research Data Alliance, and the Practices working group of the ACM's Emerging Interest Group on Reproducibility and Replicability. She sits on the board of the Roper Center for Public Opinion Research and serves on a number of advisory and task force groups working on data curation and research transparency. Prior to joining ISPS, Limor was Research Director at Northwestern University's Media Management Center and Readership Institute, and Associate Professor (clinical) at the Medill School of Journalism.
ORCID: 0000-0002-3234-1593

Zhiming Zhao received his Ph.D. in Computer Science in 2004 from the University of Amsterdam (UvA). He is an Assistant Professor in the Multiscale Network Systems (MNS) at UvA, and the technical manager of the Virtual Lab and Innovation Center (VLIC) of LifeWatch ERIC, a European research infrastructure in ecology and biodiversity science. His research focuses on innovative programming and control models for quality critical systems on programmable infrastructures such as Clouds, Edges, and Software-Defined Networking using optimization, semantic linking, blockchain, and artificial intelligence technologies. He leads the UvA effort in several EU projects including ARTICONF, CLARIFY, ENVRI-FAIR, and SWITCH.
ORCID: 0000-0002-6717-9418

Yann Le Franc, Ph.D., is the CEO and Scientific Director of e-Science Data Factory S.A.S.U., a French R&D company aiming at proposing innovative solutions for FAIR data management to accelerate growth and progress. Yann Le Franc has a Ph.D. degree in Neurosciences and Pharmacology (2004). After a postdoctoral experience in the USA, he worked on data management projects in neurosciences at the University of Antwerp (Belgium) and in the context of the International Neuroinformatics Coordinating Facility (INCF) where he developed a strong expertise in ontology design and semantic Web technologies. He then contributed to several Horizon 2020 Research Infrastructure projects (EUDAT, EOSC-Hub,. . .) as an expert on Semantic Web and ontology design. He is co-chairing the Research Data Alliance Vocabulary and Semantic Service Interest Group and the FDO Semantic Group. He is also a member of the EOSC Semantic Interoperability Task Force. He is actively involved in the FAIRification and standardization of semantic artefacts in the context of FAIRsFAIR and OntoCommons projects. In parallel, he is the technical manager of the EOSC-Pillar project for the French National Computing Center for Higher Education (CINES).
ORCID: 0000-0003-4631-418X